Adapting finite-state translation to the TransType2 project

Elsa Cubel, Jorge González, Antonio L. Lagarda, Francisco Casacuberta, Alfons Juan and Enrique Vidal
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
E-46071 València, Spain
{ecubel, jgonza, alagarda, fcn, ajuan, evidal}@iti.upv.es

Abstract

Machine translation can play an important role nowadays, helping communication between people. One of the projects in this field is TransType2 [1]. Its purpose is to develop an innovative, interactive machine translation system. TransType2 aims at facilitating the task of producing high-quality translations, and make the translation task more cost-effective for human translators.

To achieve this goal, stochastic finite-state transducers are being used. Stochastic finite-state transducers are generated by means of hybrid finite-state and statistical alignment techniques.

Viterbi parsing procedure with stochastic finite-state transducers have been adapted to take into account the source sentence to be translated and the target prefix given by the human translator.

Experiments have been carried out with a corpus of printer manuals. The first results showed that with this preliminary prototype, users can only type a 15% of the words instead the whole complete translated text.

[1] TransType2 - Computer Assisted Translation, RTD project by the European Commission under the IST Programme (IST-2001-32091).

1 Introduction

The aim of the TransType2 project (TT2) (SchlumbergerSema S.A. et al., 2001) is to develop a Computer Assisted Translation (CAT) system that will help to solve a very pressing social problem: how to meet the growing demand for high-quality translation.

The innovative solution proposed by TT2 is to embed a data driven Machine Translation (MT) engine within an interactive translation environment. In this way, the system combines the best of two paradigms: the CAT paradigm, in which the human translator ensures high-quality output, and the MT paradigm (Brown et al., 1990), in which the machine ensures significant productivity gains.

Until now, translation technology has not been able to keep pace with the demands for high-quality translation. TT2 has the ability to significantly increase translator productivity and thus has enormous commercial potential. Six different versions of the system will be developed for translation between English and French, Spanish or German. To ensure that TT2 meets the needs of translators, two professional translation agencies are in charge of evaluation of the successive prototypes.

Stochastic finite-state transducers (SFST) have proved to be adequate models for MT in limited-domain applications (Vidal, 1997; Amengual et al., 2000; Casacuberta et al., 2001). SFSTs are interesting for their simplicity and the possibility of inferring models automatically from bilingual training corpus. They allow a very efficient search

of new test data (Vidal, 1997) and make it possible to work with text-input and speech-input translation (Amengual et al., 2000).

Moreover, hybrid finite-state techniques and statistical translation techniques can be used to produce efficient SFSTs. The learning of SFST can be improved if word-aligned training pairs are used (i.e. using statistical alignment models).

This paper is concerned with the use of SFSTs for computer-assisted translation. The SFSTs have been learnt automatically from parallel corpus using an algorithm based on the statistical word-alignments and the use of n-grams (Casacuberta, 2000). The parsing (search) with SFSTs is carried out by an adaptation of the Viterbi algorithm to the framework of computer-assisted translation.

Next section is devoted to the automatic learning of SFSTs. In section 3, the search procedure for an interactive translation is presented. Experimental results are presented in section 4, where we also report some preliminary results using specialized transducers. Finally, some conclusions are shown in section 5.

2 Inference of Finite-State Transducers: GIATI

Given a source sentence s, the goal of MT is to find a target sentence $\hat{\mathbf{t}}$ that maximize:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t} \mid \mathbf{s}) = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}, \mathbf{s})$$

SFSTs are models that can be used to estimate the joint distribution $\Pr(\mathbf{t}, \mathbf{s})$ (Casacuberta, 1996). Given a SFST $\mathcal{T}$,

$$\hat{\mathbf{t}} \approx \underset{\mathbf{t}}{\operatorname{argmax}} \Pr_{\mathcal{T}}(\mathbf{t}, \mathbf{s})$$

SFST have been successfully applied into many translation tasks (Vidal, 1997; Amengual et al., 2000; Casacuberta et al., 2001). Current parsers for SFSTs produce a target sentence from a source sentence using the Viterbi algorithm.

For computer-assisted translation, the decoder must produce one (or n-) best translation prediction(s) given a source sentence and a prefix of a sentence in the target language. Given a SFST $\mathcal{T}$, a source sentence $\mathbf{s}$ and a prefix of the source sentence $\mathbf{t}_p$, the goal is to search for a suffix of the target sentence $\hat{\mathbf{t}}_s$:

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr(\mathbf{t}_s \mid \mathbf{s}, \mathbf{t}_p) \approx \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr_{\mathcal{T}}(\mathbf{t}_p \mathbf{t}_s, \mathbf{s})$$

This equation is similar to the one for general translation but in this case, the optimization is performed on a set of target suffixes rather than the set of whole target sentences.

The inference of such SFSTs is carried out by the Grammatical Inference and Alignments for Transducer Inference (GIATI) technique (the previous name of this technique was MGTI - Morphic-Generator Transducer Inference) (Casacuberta, 2000). Given a finite sample of string pairs, it works in three steps:

1. Building training strings. Each training pair is transformed into a single string from an extended alphabet to obtain a new sample of strings.
2. Inferring a (stochastic) regular grammar. Typically, smoothed n-gram is inferred from the sample of strings obtained in the previous step.
3. Transforming the inferred regular grammar into a transducer. The symbols associated to the grammar rules are transformed into input/output symbols by applying an adequate transformation, thereby transforming the grammar inferred in the previous step into a transducer.

The transformation of a parallel corpus into a string corpus is performed using statistical alignments (a function from the set of positions in the target sentence to the set of positions in the source sentence). These alignments are obtained using the GIZA software (Och and Ney, 2000; Al-Onaizan et al., 1999), which implements IBM statistical models (Brown et al., 1990; Brown et al., 1993).

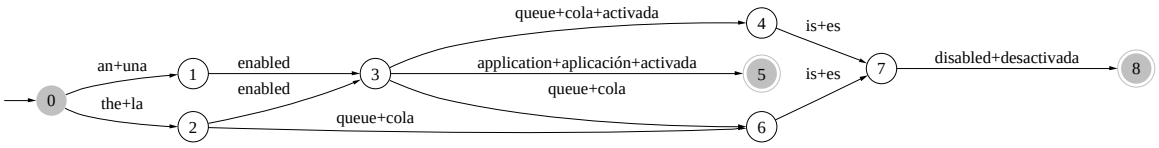

Figure 1: A non-smoothed bigram inferred from the training sentences of the example

A training string is built by assigning the corresponding aligned word from the source sentence to each word from the target sentence (Casacuberta, 2000). This assignment must not violate the order in the target sentence. Using this type of transformation from a pair of strings into a string of extended symbols, the transformation from a grammar to a finite state transducer in step 3 is straightforward.

Example 1 *An example of the GIATI application is shown.*

- *First step of the GIATI technique: From training pairs to training strings.*
  - ***Training pairs.*** *Here are some samples where the source is an English sentence and the target is a Spanish sentence.*

    an enabled queue is disabled
    → una cola activada es desactivada

    the enabled application
    → la aplicación activada

    the queue is disabled
    → la cola es desactivada

  - ***Word aligned pairs.*** *From the training pairs, this step consists on the alignment of each word from the target sentence to the corresponding word from the source sentence. For instance, English word 'an' corresponds with the Spanish word 'una'. So, in the target sentence, the word 'una' is tagged with (1) (the first word in the source sentence).*

    an enabled queue is disabled
    → una(1) cola(3) activada(2) es(4) desactivada(5)

    the enabled application
    → la(1) aplicación(3) activada(2)

    the queue is disabled
    → la(1) cola(2) es(3) desactivada(4)

  - ***Training sentences.*** *Sometimes, the alignment produces a violation of the sequential order of the words in the target sentence. For example, in the first sentence, if the Spanish word 'cola' is assigned to the English word 'queue' and the Spanish word 'activada' to the English word 'enabled', it implies a reordering of the words 'cola' and 'activada'.*
    In order to prevent this problem, the output word is assigned to the first input word that does not violate the output order. For instance, in the first sentence, 'enabled' should be assigned to 'activada', but as this implies a reordering of the words, 'enabled' is assigned to ' ' and 'queue' is assigned to 'cola activada' because 'queue' is the next word that does not violate the output order.

    an**+una* ***enabled
    queue**+cola+activada* ***is**+es* ***disabled
    +desactivada

    the** +la* ***enabled
    ***application**+aplicación+activada*

    ***the**+la* ***queue**+cola* ***is**+es*
    ***disabled**+desactivada*

- *Second step of the GIATI technique: From training strings to grammars (n-grams). Here, a (stochastic) regular grammar is inferred (Figure 1) from the strings that are built from first step.*

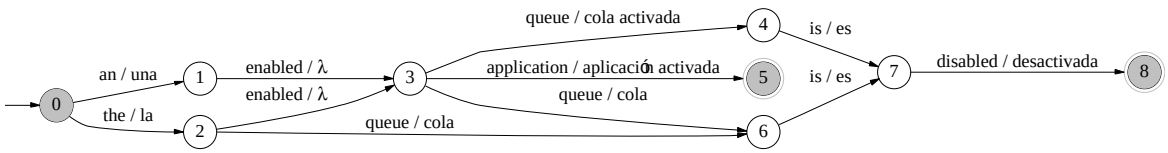

Figure 2: The Finite-State Transducer that is obtained from the bigram of Figure 1

The n-grams are particular cases of stochastic regular grammars that can be inferred from training samples.

- *Third step of the GIATI technique: From grammars to transducers.*

  From the grammar obtained in the second step, a finite-state transducer is obtained. It is shown in Figure 2.

  Each transition label is transformed into a source/target pair. The source part is composed of the source word in the transition label and the target part is the sequence of target words in the transition label.

An interesting feature of the GIATI method is the fact that all the techniques which are known for n-gram smoothing are directly applicable in the second step of the method.

3 Search in an interactive manner

The translation procedure of the GIATI system is based on Viterbi search (Viterbi, 1967; Picó and Casacuberta, 2001) for the optimal path in a finite-state network. The translation of a source sentence is built by concatenating the target strings of the successive transitions that compose the optimal path.

There are two steps: 1) Searching for the most probable path of states that deals with the source sentence and the prefix given by the user; and 2) from the state achieved by the optimal path in step 1, searching for the optimal path that deals with the rest of source sentence.

Example 2 *An example of this procedure is presented. It is based on the finite-state transducer obtained in Figure 2.*

In this case, the source English sentence is 'the enabled queue is disabled'. Let us suppose that the Viterbi search is applied, the most probable path corresponds to the Spanish sentence 'la cola es desactivada' (Left side of Figure 3).

By contrast, let us suppose that in Viterbi for computer-assisted translation, the prefix introduced is 'la cola activada'. This prefix does not correspond with the beginning of the most probable sentence found by traditional Viterbi. So, the new Viterbi (Right side of Figure 3) is forced to find a target sentence beginning with the prefix. Once the prefix is forced, the rest of the sentence, that is 'es desactivada', is searched in the traditional method.

This technique entails some problems. One of them appears when the target word suggested by the user is in the target vocabulary but the prefix is not in the transducer. The current solution for this case is to apply smoothing. Besides, if the suggested word is not in the target vocabulary, it is associated to the UNK (unknown) symbol, and a new entry to this word is added to the vocabulary.

To reduce the computational cost of the search, the beam-search technique has been implemented. A future version of the search will deal with target word graphs.

4 Experimental results

A TT2 interactive prototype, which uses the searching techniques presented in the previous sections, has been implemented.

Example 3 *An example of TT2 interactive search process is analyzed in detail for English-to-Spanish translation of printer manuals.*

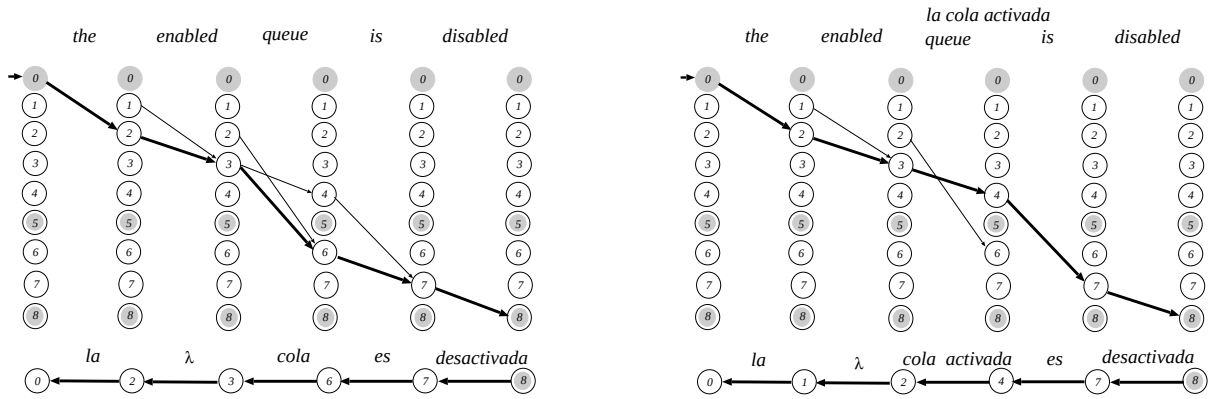

Figure 3: Viterbi search for computer-assisted translation

Source sentence: *It also contains a section to help users of previous software versions adapt more quickly to the new software.*

Hypothesis 0: *Se se para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software.*

Prefix 0: ***También***

Hypothesis 1: ***También*** *se ofrece una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software.*

Prefix 1: ***También contiene***

Hypothesis 2: ***También contiene*** *una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software.*

Final hypothesis: *También contiene una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software.*

From the source sentence, the system provides a target sentence (Hypothesis 0) as its best hypothesis. This hypothesis is clearly incorrect and the user changes the prefix 'Se' *by* 'También' *(Prefix 0). Now, the system searches for the most probable path in the transducer beginning with the prefix. The new hypothesis is still not considered to be completely acceptable by the user, who amends it with a new, longer prefix,* 'También contiene'*. The new hypothesis, that has these two words as a prefix, is already judged satisfactory by the user who accepts it as the final result.*

The prototype has been applied in the TT2 project for the Xerox task. It involves the translation of technical Xerox manuals from English to Spanish. In the training corpus, all numbers have been substituted by a single category. The data used for training a test set are shown in the Table 1.

Table 1: Features of the Xerox Corpus

Data		English	Spanish
Training	Sentence pairs	45,493	
	Running words	517,534	575,776
	Vocabulary	9,795	12,211
	Trigram test-set perplexity	29.1	24.8
Test	Sentences	500	
	Running words	5,719	6,425
	Running characters	–	32,165
	Average chars. per word	–	5.5

The assessment of the prototype has been carried out using three measures:

1. *Translation Word Error Rate* (TWER). Edit distance between the output final sentence of the translator and a reference translation.
2. *Number of Word Corrections* (NWC). Number of user interactions that are necessary to achieve the reference targets divided by the number of running words. In each interaction only one wrong word is changed.

3. *Key-Stroke Ratio* (KSR). Number of key-strokes that are necessary to achieve the reference targets divided by the number of running characters. In each interaction only one character is changed (This measure has been estimated from NWC by assuming that only 50% of characters in the wrong word must be corrected).

The achieved results are shown in the Table 2. There are different values of the n-grams in the GIATI technique.

Table 2: Results for the Xerox Corpus. TWER is the Translation Word Error Rate (%), NWC is the Number of Word Corrections (%) and KSR is the Estimated Key-Stroke Ratio (%). The first column corresponds to the n-gram used in the GIATI technique

n-gram	TWER	NWC	KSR	States	Transitions
2	36.4	30.3	40.1	62,476	501,237
3	33.4	15.0	32.5	250,620	951,875
4	32.9	15.0	32.5	513,114	1,483,357
5	32.9	14.7	32.3	519,609	1,501,070

Best results were achieved when four-grams were used in the GIATI technique. However, with trigrams the results are very similar and the size of the transducer is clearly lower. It is interesting to note that each interaction improves the translation so that NWC score is match smaller than the TWER score.

Apart from the results obtained with the above single-model approach, we have also obtained some preliminary results with a new, multiple-model approach. The basic idea is to learn a text classifier from source sentences which is used for both, training and testing. On the one hand, the classifier is used during training to divide the data into homogeneous groups from which specialized transducers are learned. During testing, on the other hand, the classifier selects the most likely specialized transducer for each source sentence so as to get specialized translations. So far the only (unsupervised) text classification model we have tried is a *mixture of multinomial distributions*, with each mixture component playing the role of a different class (see (Juan and Vidal, 2002) for details on a similar mixture model). Using this mixture of multinomials model, we got the results shown in Figure 4, in which the TWER and (average) sentence translation time (TIME) are given as a function of the number of mixture components.

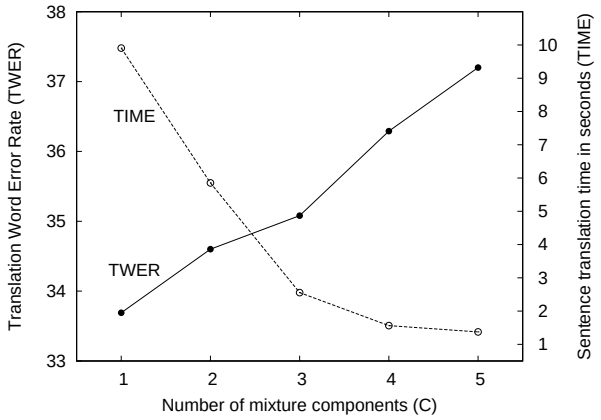

Figure 4: Results obtained with the multiple-model approach and a mixture of multinomials classification model. The plot shows the translation word error rate (TWER) and average sentence translation time (TIME) as a function of the number of mixture components

The results given in Figure 4 are not as good as expected since the best TWER is obtained with a single-component mixture, that is, with the conventional (single-class) approach. Moreover, the TWER degrades as the number of mixture components increases so, possibly, our multiple-model approach suffers from lack of training data (the more mixture components are used, the less source sentences we have, on average, to train each specialized transducer). We have not investigated this further. For the time being, we are somehow satisfied with the fact that our new approach enables an easy way to reduce the computing time alloted to sentence translation. This is particularly important in the TT2 project and, for instance, it could be said that our 2-component mixture model is preferable to the conventional (single-component) model because it approximately halves the translation time without too much degradation in translation quality.

5 Conclusions and future work

Finite-state transducers can be used for computed-assisted translation. These models can be

learnt from parallel corpus, but the number of states/transitions can be too high. In this case, finite-state models for specific topics could be used.

Specialized transducers have been trained using multinomial mixture modelling, but no improvements have been found so far in terms of TWER. Nevertheless, some computational advantages can be obtained without sacrificing too much translation quality.

Some improvements of this technique will be the production of N-best hypothesis (rather than one hypothesis) and the production of short sequences of words (rather than whole suffixes). Another future work will be the introduction of morpho-syntactic information and/or bilingual categories in the finite-state transducers.

Acknowledgements

The authors would like to thank the reasearchers involved in the TT2 project who have developed the methodologies that are presented in this paper.

This work has been supported by the European Union under the IST Programme (IST-2001-32091).

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz J. Och, David Purdy, Noah Smith, and David Yarowsky. 1999. Statistical machine translation.

Juan C. Amengual, José M. Benedí, Asunción Castano, Antonio Castellanos, Víctor M. Jiménez, David Llorens, Andrés Marzal, Moisés Pastor, Federico Prat, Enrique Vidal, and Juan M. Vilar. 2000. The EuTrans-I speech translation system. *Machine Translation*, 15:75–103.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

Francisco Casacuberta, David Llorens, Carlos Martínez, Sirko Molau, Francisco Nevado, Hermann Ney, Moisés Pastor, David Picó, Alberto Sanchis, Enrique Vidal, and Juan M. Vilar. 2001. Speech-to-speech translation based on finite-state transducers. In *International Conference on Acoustic, Speech and Signal Processing*, volume 1. IEEE Press, April.

Francisco Casacuberta. 1996. Maximum mutual information and conditional maximum likelihood estimations of stochastic syntax-directed translation schemes. In L. Miclet and C. de la Higuera, editors, *Grammatical Inference: Learning Syntax from Sentences*, volume 1147 of *Lecture Notes in Artificial Intelligence*, pages 282–291. Springer-Verlag.

Francisco Casacuberta. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In A.L. Oliveira, editor, *Grammatical Inference: Algorithms and Applications*, volume 1891 of *Lecture Notes in Computer Science*, pages 1–14. Springer-Verlag. 5th International Colloquium Grammatical Inference -ICGI2000-. Lisboa. Portugal.

Alfons Juan and Enrique Vidal. 2002. On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705–2710, December.

Franz J. Och and Hermann Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hongkong, China, October.

David Picó and Francisco Casacuberta. 2001. Some statistical-estimation methods for stochastic finite-state transducers. *Machine Learning*, 44:121–142, July-August.

SchlumbergerSema S.A., Intituto Tecnológico de Informática, Rheinisch Westfälische Technische Hochschule Aachen Lehrstul für Informatik VI, Recherche Appliquée en Linguistique Informatique Laboratory University of Montreal, Celer Soluciones, Société Gamma, and Xerox Research Centre Europe. 2001. TT2. TransType2 - computer assisted translation. Project technical annex.

Enrique Vidal. 1997. Finite-state speech-to-speech translation. In *Int. Conf. on Acoustics Speech and Signal Processing (ICASSP-97), proc., Vol.1*, pages 111–114, Munich.

Andrew Viterbi. 1967. Error bounds for convolutional codes and a asymtotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.